DOCUMENT RESUME

ED 348 382 TM 018 756

AUTHOR Shepard, Lorrie A.
TITLE Will National Tests Improve Student Learning?
INSTITUTION Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.; Colorado Univ., Boulder.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO CSE-TR-342
PUB DATE Apr 92
CONTRACT R117G10027
NOTE 23p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Advisory Committees; Cost Effectiveness; *Educational Improvement; Elementary Secondary Education; *Learning; *National Competency Tests; *Standardized Tests; Student Evaluation; Test Construction; Testing Problems; *Test Use
IDENTIFIERS *National Education Goals 1990; National Standards; Performance Based Evaluation; Standard Setting

ABSTRACT

Claims that national tests will improve student learning are explored, asking whether national examinations will ensure high-quality instruction and greater student learning and whether tests developed to meet urgent political deadlines will retain essential features of authentic curriculum-driven assessments. Part I presents research evidence on the negative effects of standardized testing, such as the effects of high stakes testing on scores, the curriculum, and instruction. The National Education Goals Panel's (NEGP's) version of national examinations is presented in Part II, with attention to their proposals intended to forestall the negative effects of traditional tests. Part III identifies curricular and technical problems that must be resolved before the NEGP's vision can be achieved. These include: (1) development of world class rather than lowest common denominator standards; (2) development of incorruptible performance tasks; (3) teacher training in curriculum and instruction; (4) high standards for all students without reinstitution of tracking; and (5) cost. If tests are developed before these problems are resolved, new tests are likely to have the same pernicious effects as the old. There is a 32-item list of references. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED348382

CRESST

National Center for Research on Evaluation, Standards, and Student Testing

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Will National Tests Improve Student Learning?

CSE Technical Report 342

Lorrie A. Shepard

University of Colorado, Boulder

April 1992

- UCLA Center for the Study of Evaluation

in collaboration with:

- University of Colorado
- NORC, University of Chicago
- LRDC, University of Pittsburgh
- The RAND Corporation

TM018756

BEST COPY AVAILABLE

Will National Tests Improve Student Learning?

CSE Technical Report 342

Lorrie A. Shepard

University of Colorado, Boulder

April 1992

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
UCLA Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

The work reported herein was partially supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

Will National Tests Improve Student Learning?*

Lorrie A. Shepard

University of Colorado, Boulder

The current frenzied interest in testing is motivated by a desire to improve public education. Policy makers believe that by setting standards and measuring their attainment teachers will be exhorted to teach better and students to learn more. This idea of test-leveraged reform is not new. The same logic motivated minimum competency testing in the 1970s and the educational reform movement of the 1980s. But previous efforts at test-driven reform failed. The authors of *A Nation At Risk*, in 1983, specifically rejected "minimum competency" examinations (then required in 37 states) because "the 'minimum' tends to become the 'maximum,' thus lowering educational standards for all" (p. 20). Now, less than a decade later, we have evidence that the standardized testing programs instituted in response to *A Nation at Risk* persisted in limiting what students learned. Nationally, basic skills test scores have increased at the expense of higher-order thinking and problem solving.

The most savvy proponents of a new national testing effort promise that this time things will be different. Better tests—performance measures aimed at assessing higher-order learning goals—will ensure student learning by redirecting instruction toward more challenging content. Thus, this decade's high-stakes accountability tests are expected to play the same role as before in forcing school reform, but the effects will be different because of a fundamental change in the character of the new assessments.

The purpose of this paper is to analyze these claims. Will national examinations ensure high-quality instruction and greater student learning? Will tests developed in the short-term to meet urgent political deadlines retain the essential features of authentic, curriculum-driven assessments? The

---

* Paper presented at the American Educational Research Association Public Interest Invitational Conference, Accountability as a State Reform Instrument: Impact on Teaching, Learning, and Minority Issues and Incentives for Improvement, Washington, DC, June 5, 1991.

paper is organized in three parts. Part I is a summary of research evidence on the negative effects of standardized testing. In Part II, the National Education Goals Panel's (NEGP) vision of national examinations is presented with particular attention to the features of their proposal intended to forestall the negative effects experienced with traditional tests. In Part III, curricular and technical problems are identified that have to be resolved before the vision of the NEGP can be achieved. If national tests are installed before these problems are solved, what is the likelihood that testing will have the same negative effects as described in Part I?

Part I: Research Evidence on the Negative Effects of High-Stakes Standardized Testing

In previous debates about testing, claims and counterclaims about the effects of externally mandated tests have been largely rhetorical. Proponents of measurement-driven instruction (MDI) argued, in the 1980s, that high-stakes tests would set clear targets thus assuring that teachers would focus greater attention on essential basic skills (Popham, Cruse, Rankin, Sandifer, & Williams, 1985). Critics countered that measurement-driven instruction distorts the curriculum, "MDI fragments [curriculum], narrows it, deflects it, trivializes it, and causes it to stagnate" (Bracey, 1987). Each side argued theoretically and from limited observations but without systematic proof of these assertions.

Today there is a body of research evidence that informs the debate about the effects of externally mandated standardized testing. Key findings are summarized below. To the extent that tests rely on multiple-choice formats and focus on basic skills, these findings apply to both commercial norm-referenced tests and to criterion-referenced tests developed by some states and districts.

1. When test results are given high-stakes by political pressure and media attention, scores can become inflated, thus giving a false impression of student achievement.

The Cannell report released in 1987, debunking the claim that all 50 states are above average, drew public attention for the first time to the possibility that euphoric tests score gains might be fraudulent. In a systematic follow-up of

the Cannell report, Linn, Graue, and Sanders (1990) concluded that achievement trends reported by states and districts on the basis of standardized tests appear to be increasing more dramatically than can be corroborated by parallel comparisons on secure tests administered by the National Assessment. Recently we undertook a study to evaluate the credibility of publicly reported standardized test scores in high-stakes school districts by administering unfamiliar tests covering the same content. Preliminary results in one large district indicate that students know dramatically less reading and math content when given an independent assessment than they appear to know on their routinely administered standardized tests (Koretz, Linn, Dunbar, & Shepard, 1991).

We do not believe, for the most part, that spuriously high test scores are the result of blatant cheating like practicing on the test beforehand or giving students answers. Rather, the widespread inflation of standardized test scores is more likely due to teaching the test in a more general sense, that is, giving students extensive practice with the kinds of questions that appear on the test in precisely the same format as found on the test. Experimental studies have shown that students who receive narrowly focused coaching on one type of test question are often unable to answer correctly if the same skill is tested in even a slightly different way (Shepard, 1988). Thus it is possible to raise test scores without increasing learning.

2. High-stakes tests narrow the curriculum. Tested content is taught to the exclusion of non-tested content.

Although critics may have feared originally that testing would take instructional time away from "frills" such as art and citizenship, the evidence now shows that social studies and science are neglected because of the importance of raising test scores in basic skills (Darling-Hammond & Wise, 1985, Rottenberg & Smith, 1990, Shepard, 1991). Even in reading, math, and language mechanics, instruction is focused only on skills covered on the test. In an extensive 18-month observational study, for example, Rottenberg and Smith (1990) found that because of external tests elementary teachers had given up on reading real books, writing, and long-term projects and were filling all available time with word recognition, recognition of errors in spelling, language usage, punctuation, and arithmetic operations.

Available research also suggests two disturbing trends. The degree to which testing distorts curriculum can be predicted from (a) the extent of political pressure (i.e., the greater the perceived stakes of test results, the greater the narrowing of curriculum), and (b) the socioeconomic level of the school district (i.e., the poorer the school or district, the more time is given to teaching to tests).

3. High-stakes testing misdirects instruction even for the basic skills.

Under pressure classroom instruction is increasingly dominated by tasks that resemble tests. Test-driven curriculum encourages teaching of skills in isolation, and the test-like format of materials elicits different cognitive processes than if teachers had addressed intended learning goals directly. For example, students are asked to read artificially short texts and to recognize one right answer rather than invent either their own questions or possible answers. Even in the early grades, they practice finding mistakes rather than writing, and learn to guess by eliminating wrong answer choices.

In a recent survey 68% of teachers in high-stakes districts reported that they spent time "regularly throughout the school year" preparing students to take tests rather than just the few weeks before testing (Shepard, 1991). Individual teachers elaborated their answers with comments such as the following, "Time I would like to have spent using the Whole Language approach was used up in practice tests and district practice materials. I felt pressure from my principal to complete all practice materials, because she feels that is the road to success." "Critical thinking skills are basically non-existent in our children because of drill and practice for [three different tests]." "The students receive little hands-on learning in place of drill and specific skills teaching."

4. The kind of drill-and-practice instruction that tests reinforce is based on outmoded learning theory. Rather than improve learning, it actually denies students opportunities to develop thinking and problem-solving skills.

As documented by Resnick and Resnick (in press), traditional multiple-choice achievement tests require students to demonstrate decontextualized, atomistic skills. The *decomposability* and *decontextualization* assumptions underlying the development of traditional tests were based on early learning theories (associationism and behaviorism) that have since been rejected by

contemporary research in cognitive psychology. According to the old theories, complex, higher-order skills had to be acquired bit-by-bit by breaking learning down into constituent, prerequisite skills. It was assumed incorrectly that after basic skills had been learned by rote, they could be assembled mechanistically into complex understandings and insight. In contrast, evidence from cognitive psychology shows that all learning requires thinking and active construction by the learner of evolving mental models or schemas.

When teachers teach to traditional tests by providing daily skill instruction in formats that closely resemble tests, their instructional practices are not just ineffective but detrimental. By following a theory that postpones the development of higher-order thinking skills until after the basic have been mastered, teachers deny learning opportunities in two ways. First, learning isolated facts and skills becomes more difficult because without context there is no meaningful way to chunk or organize information and make it easy to remember. Second, learning decontextualized skills means that later application of skills to solve real-world problems becomes a separate and difficult learning hurdle.

Low scoring students are especially hurt by test-driven instruction because they are consigned full-time to deadly boring drill. Having failed to master the basics, they never get to go on to the "good stuff" that might have made school learning interesting. The negative consequences of instruction that perseverates on prerequisite skills can be seen in the attitudinal domain as well. Borko and Eisenhart (1986) found, for example, that children who had been assigned to reading groups on the basis of standardized test results had fundamentally different views of what reading was about. Only high group students mentioned understanding or meaning as the purpose of reading whereas poor readers thought that their goal was fluency, that is, to read orally without decoding errors. And indeed, classroom observations revealed that only the students in the high group received instruction aimed at developing their comprehension skills.

5. Because of the pressure on test scores, more hard-to-teach children are rejected by the system. There is a direct correspondence between accountability pressure and the number of children denied kindergarten entrance, assigned to two-year kindergarten programs, referred to special education, made to repeat a grade, or who drop out of school.

Observational and teacher-interview studies have found a connection between the importance of standardized tests in a district or school and the tendency to screen children out of school or retain children in kindergarten (California State Department of Education, 1988; Cunningham, 1989; Hatch & Freeman, 1988; Shepard & Smith, 1988). These practices are particularly worrisome because controlled studies do not show any improvement in achievement from two-year kindergarten programs and do show social-emotional harm (Shepard, 1989). Denying school entry to some children on the basis of readiness tests (for fear that they will not do well on standardized achievement tests in later grades) is problematic because the process screens out a disproportionate number of poor and minority children who most need access to public education.

Of course, there is an explicit link between standardized test results and grade retention in situations where tests are used to make pass-fail decisions. These practices are intended to ensure student learning, but the overwhelming weight of evidence is that grade retention lowers achievement in subsequent years, harms self esteem, and substantially increases the risk of dropping out (Shepard & Smith, 1989). Likewise, competency testing affects the dropout rate. Despite the belief among most educators that minimum competency graduation tests are such low-level barriers that they do not seriously affect students' progress, students themselves are discouraged by failing these exams and are therefore less likely to continue in school (Catterall, 1988).

6. The dictates of externally mandated tests reduce both the professional knowledge and status of teachers.

Hatch and Freeman (1988) found that teachers were themselves victims of instructional decisions dictated by accountability pressures. Although teachers were the direct agents of worksheet, skill-based programs, 67% reported considerable distress because of the discord between the instructional methods they were forced to adopt and their own training and beliefs about children's learning needs. Hatch and Freeman likened the systematic stress and role conflict that they encountered among kindergarten teachers to earlier research findings on beginning teachers who are denied professional autonomy (Walberg, 1970). The likelihood is that the most intellectually able

teachers and those with the greatest sense of professional identity will leave teaching because of the conflict, and those who remain are likely to protect themselves from the dissonance by caring less about their teaching and their students.

Test-leveraged educational reform is a quintessential example of bureaucratic accountability. As explained by Darling-Hammond (1988), the very conception of bureaucratic accountability is intended to remove control from the judgments of individual teachers, hence the notion of "teacher-proof" curricula, etc. Teachers are not expected to be experts in child development or pedagogy but to implement externally mandated curricula, textbooks, tests, and promotional standards as specified. Bureaucratic accountability (which Darling-Hammond contrasts with professional accountability) begets exactly what it assumes, namely less skilled professionals. For example, in an extensive field study, McNeil (1988) found that two reforms, proficiency-based testing of students and a checklist assessment of teacher behaviors, actually reinforced the dynamics of low-quality, mechanistic instruction. Similarly, Rottenberg and Smith (1990) found that teachers were both degraded and deskilled by a high-stakes testing environment. Teachers felt ashamed and embarrassed by low scores even when they recognized the influence of socioeconomic factors on school rankings and the mismatch of the test content to instructional goals. And teachers were deskilled because their decisions to align instruction to the tests impoverished their teaching repertoires and ultimately limited their own conceptions of what should be taught.

Part II. National Examinations Envisioned By the National Education Goals Panel

The debilitating negative effects of standardized achievement tests have been recognized by an increasingly large circle of curriculum and measurement specialists, educators, and policy makers. Alternatives have been proposed under a variety of different names. Authentic assessments, direct assessments, and performance assessments share the common idea that measurement tasks should be instances of complex, criterion performances that directly represent the ultimate goals of education. Thus, practice on such tasks would lead instruction in a positive rather than a

negative direction. Authentic assessment tasks contrast with the kinds of questions asked on traditional standardized tests which were only proxies or indicators of real achievement. Although multiple-choice test questions might indeed correlate highly with performance measures under circumstances where neither measure is used for accountability, it is the proximate nature of traditional tests that makes them vulnerable to inflation and to misguided instruction when they are used in high-stakes contexts.

Wiggins (1989) used the term "authentic" assessment to convey the idea that assessments should elicit the actual performances that we want students to be good at. Authentic assessment tasks should be complex and embedded in local contexts; they should test for mastery of concepts and insights central to the discipline; and they should allow for demonstration of students' habits of mind, for example, how they frame a question as well as how they solve it. In the same vein, Frederiksen and Collins (1989) argued for direct assessments that would involve students in extended tasks requiring the demonstration of the cognitive skills intended as the goals of instruction. Frederiksen and Collins wanted to use the power of important assessments to direct student and teacher effort. In addition, the standards and scoring rubrics of these assessments would become "the medium for communicating to teachers and students the critical traits to look for in good writing, good historical analysis, and good problem solving" (p. 29). Resnick and Resnick (in press), who have been the most directly involved in the development of the NEGP's recommendations, also argued for alternative assessments that would embody the goals of the thinking curriculum and thereby serve as a positive catalyst for change whenever teachers did the "natural thing," that is, prepared their students to do well on high-stakes accountability measures.

All of these proposals assume that because of the nature of the assessment tasks "teaching to the test" will have only salutary effects on the quality of instruction and student learning. Although these authors are primarily interested in the effects of performance assessment on instructional change, it is also the case that direct measures (if incorruptible) should be more valid measures of student achievement.

The Resource Group of the National Education Goals Panel (1991) responsible for Goal Three has proposed a model for an end-of-decade national examination system based on the principles of authentic, performance

assessment. Although the Resource Group also explained the need to preserve the National Assessment of Education Progress (NAEP) as a monitoring system, the focus here is on their proposal for a second system to examine individual students, which they described as follows:

> The fundamental reason for introducing a system in which individual students are examined is the belief that such examinations can provide focused targets for study and instruction, thus raising achievement levels in the Nation. To serve this purpose, examinations must function as an integral part of the educational system, capable of setting a clear standard toward which students and teachers can work. This goal dictates the kinds of assessments that will be needed. They must be assessments focused on high levels of achievement (thinking and reasoning, not routinized skills), assessments tied to curriculum goals or frameworks, assessments designed to be studied for and taught to. No broad assessment system that meets all three of these criteria and is designed for our entire school population is available today. Building such systems will require substantial effort and resources over the decade. (pp. 46-47)

A system of national examinations as conceived by the NEGP Resource Group would have high stakes but would avoid the problems of narrowed curriculum and rote instruction by developing examinations based on ambitious, world-class standards. If the examinations were curriculum or syllabus driven, there would be in theory no distinction between practicing on tasks mimicking the test and good instruction. It is also the intention of the Resource Group that the examinations not lead to sorting and segregation of students. They suggest, for example, that "consideration should be given to permitting students to accumulate examination credits over several years rather than sitting for a single pass-or-fail test. This method could accommodate different rates of learning, yet hold all students to a high standard of achievement" (p. 52). Although the Resource Group did not address explicitly how a national examination system would be respectful of local teacher autonomy, elsewhere in their proposal they promote this as a value in judging the merits of state reform efforts, asking for example, "Has the State created a structure within which teachers and other local school professionals are given the freedom and responsibility to best figure out how to achieve the goals and targets established at the State level?" (p. 50).

From their work, Resnick and Resnick (in press) induced three principles as guidelines for accountability assessments: (a) You get what you assess. (b) You do not get what you do not assess. (c) Build assessments toward which you want educators to teach. In effect, proponents of performance assessments for accountability purposes have adopted the measurement-driven instruction model but with new content. The mechanism for leveraging reform is the same, but the intention is to forestall negative effects by building better tests.

Part III: Problems To Be Resolved Before Beneficial National Examinations Can Be Instituted Rather Than Just More Negative Tests

The vision of curriculum-driven examinations offered by the National Education Goals Panel is inspired. However, we do not at present have the technical, curricular, or political know-how to install such a system—at least not on so large a scale. As acknowledged by the Panel, "No broad assessment system that meets [their] criteria is available today" (p. 47). Advanced Placement examinations which are often cited as examples of challenging, curriculum-driven measures are not intended for a universal population of students. Even for elite groups of students, there are many curriculum experts who challenge the wisdom of the AP curricula that often trade breadth over depth in a subject area. Moreover, there is no evidence available about what would happen to the quality of instruction if all high-school teachers, not just those who volunteered, were required to teach to the AP curricula. Another example of an existing advanced test is the New York Regents' Examination. Yet Schoenfeld (in press) found that an award-winning teacher had taught students to memorize the steps in all 12 of the geometry proofs that might appear on the exam rather than teaching them the underlying concepts. (Note that because the proofs required students to construct figures, they qualify superficially at least as performance measures.)

Just because we have not yet solved the problems of standards development, performance assessment task development, teacher training, and scoring, does not mean that we should not embark on the venture. However, it should be clear that creation of a national examination system should be viewed as an immense developmental effort, one that will require on-

going evaluation, revision, and incremental improvement. The focus of evaluation efforts will be to determine not just the technical adequacy of the measures but whether these examinations have the desired effects on teaching and learning. (Remember that only a few short years ago the measurement specialists and politicians who installed high-stakes basic skills tests were absolutely convinced that they would improve public education. It was not until the research evidence in Part I began to be collected that inaccurate rhetorical claims could be deflected.)

Once it is clear how great a development effort is needed to create examinations that will improve rather than cheapen instruction, two additional points follow. First, it should be obvious that tests administered in the short term are likely to perpetuate many of the negative effects of current tests. Second, if we don't have answers to a myriad of psychometric and instructional questions, it is better to have multiple development efforts rather than one. If there were only one test, we would not be able to evaluate its effects nor study how variations in test features affect instruction. For example, if only basic skills tests had been administered in the last decade we would not have been able to establish the effect of basic skills emphases on higher-order skills. Nor would we have been able to evaluate spurious gains in standardized test scores, if we had not had an independent assessment of trends from the NAEP (Linn et al., 1990).

The NEGP Resource Group understood the need for multiple, concurrent development efforts and recommended a "cluster" model whereby states or clusters of states would develop shared curriculum frameworks and exams. The exams would then be calibrated to a national standard. However, the President's proposal for American Achievement Tests in *America 2000* (U.S. Department of Education, 1991) departs sharply from the recommendations of the National Education Goals Panel, despite the assertion that "we expect to follow the Panel's lead in developing the New World Standards and the American Achievement Tests" (p. 32).

The only detail available on the President's version of new tests appears in the question and answer section of *America 2000*:

Q. When will the new tests be ready?

A. In 1994, we will have available a system of high quality individual tests, at least in reading, writing and mathematics—education's traditional "three R's"—for states and localities that want them. Because the new American Achievement Tests probably cannot be perfected that quickly, we will ask Congress to authorize the rapid deployment of an individual version of tests used by the existing National Assessment of Educational Progress. (U.S. Department of Education, 1991, p. 32)

In what follows, several problems are presented briefly that must be resolved before a beneficial national examination system can be fully established. Consider for each problem what the prospects are for successful resolution depending upon whether the proposal offered by the President or that of the National Education Goals Panel is adopted.

1. Development of world-class rather than lowest-common-denominator standards.

The more people who come to the table to arrive at consensus over the content of new tests, the more likely it is that ambitious, challenging content will be negotiated out. Although there will be pressure to ensure that these new tests go beyond the basics somewhat, it will be politically impossible for the test frameworks to advance very far beyond existing state curriculum frameworks, many of which reflect close alignment to their own standardized tests. Even the current National Assessment grade four mathematics framework, which was developed through a consensus process for state-by-state comparisons, has been described by experts as an improvement over the previous NAEP framework but one that falls short of the standards set by the National Council of Teachers of Mathematics (1989). In addition to political motivations that would make states unwilling to agree to assessments that go beyond state curricula, there are legal constraints against giving a test that matters to students who have not had the opportunity to learn what the test measures (Debra P. v. Turlington, 1983).

If 50 states have to agree on standards for assessment content, and if the development of the new tests is turned over to a commercial test publisher to be distributed for profit as currently proposed, the situation is ripe for exactly the

same watering down of content that has plagued the textbook industry (Tyson-Bernstein, 1988). In contrast, there are several states that are out in front in the development of their own alternative assessments: Arizona, California, Connecticut, Illinois, Kentucky, Michigan, and Vermont. How much more likely is it that desired instructional reforms will occur if these states and their neighbors each work on developing and improving their assessments than if a single national test were imposed?

2. Development of incorruptible performance tasks.

Although in theory authentic assessment tasks should elicit good teaching if taught to, this will not necessarily be the case in practice. Just as geometry proofs could be memorized, thus undercutting the goal of teaching for mathematical understanding, so most performance measures would become invalid if practice leads to apparent proficiency on specific tasks but does not generalize to other similar tasks.

For the most part it is not known to what extent the promises of authentic, performance assessment will hold true when implemented for large-scale accountability purposes. The effects of these types of instruments on instruction and their credibility in high-stakes environments can only be evaluated after they have been implemented applying the same kinds of research designs used to uncover the effects of standardized tests.

3. Teacher training in curriculum and instruction.

Some would argue that it is better to have poorly trained teachers teaching to high-level tests than to low-level tests. In other words, if teachers lack sufficient pedagogical and content knowledge to make their own instructional decisions, it is preferable that they mimic performance tests rather than multiple-choice tests. However, this lesser-of-evils choice hardly warrants the tremendous investment necessary to develop the NEGP's examinations. For the kind of transformation envisioned by National Goals Panel to take place, it is essential that the examinations follow from curriculum revisions and that teachers be full partners in making instructional changes.

External reforms imposed on teachers without adequate participation and training will negatively affect teachers' professional identities as experienced with standardized tests, and will only superficially affect day-to-day instructional decisions. For example, in a recent study of the new New York

fourth-grade hands-on science assessment, Bauer, Mathison, Merriam, and Toms (1990) found that teachers used more hands-on tasks in their instruction than they had before. However, teachers did not increase the amount of time spent on science nor did they use the syllabus to plan their lessons. Thus it appeared that the effect of this performance assessment might have been slightly positive but still superficial. Teachers had not developed the understandings necessary to transform their instruction and make the new kinds of tasks an integral part of that instruction.

Again it is more likely that the cluster model of test development pursued over a longer period of time will involve more teachers in the development of standards, reform of curriculum, development and tryout of assessment tasks, participation in scoring, and participation in further revisions of the assessment system. If one studies the participatory nature of the Vermont portfolio scoring efforts (an impressive scheme but a potential logistical nightmare even in a state as small as Vermont), it becomes clear how much geography and sheer numbers will influence the ability of an assessment program to stay closely tied to instruction.

4. High standards for all students without reinstitution of tracking.

American schools do not have experience with high-stakes tests used to make decisions about individual students that have not led to tracking and sorting decisions. It is easy to foresee that challenging tests, especially those administered in high school, will lead to tracking if only those students who are thought to be capable of the material are admitted to the test-preparation courses. Such is the case now with participation in Advanced Placement courses as well as for sorting at the bottom end where students are assigned to special remedial courses to pass minimum competency exams.

Questions of equity remain to be addressed for these new examinations. Again the question is not just whether the exams measure fairly but also how they control educational opportunity. It is only conceivable that the intention of the Goals Panel to have all students meet high standards can be achieved if instruction changes profoundly. This is literally not something that today's educational system can produce. It seems more likely that the kinds of changes needed to include a large proportion of students in thinking and

problem solving could occur over the longer term, if assessment efforts are embedded in systemic curricular changes.

5. Cost.

Assessments that do not rely on the efficiency of multiple-choice tests scored by optical scanning equipment are appreciably more expensive. For example, the recently administered fourth-grade NAEP mathematics test cost $138 per pupil (in the sample) to administer and score. In previous discussions when the issue was to redirect large-scale district and state level testing aimed at producing aggregate scores, policy makers could be encouraged to attain the benefits of performance assessments by using sampling procedures that would literally trade quantity for quality in assessment data (Resnick & Resnick, in press; Shepard, 1989). Now, however, the demand is for accountability measures administered to individual students so that their progress and accomplishments can be assessed.

There are two ways to solve the cost problem when every pupil is to be tested. One way is to resort to the efficiencies of traditional testing procedures, which is a very real danger with the President's short-term tests. The other strategy is to involve teachers in administering and scoring assessments (without pay) as part of their normal assignment. Early experiences in California, Connecticut and Vermont suggest that involving teachers in this way serves important staff development functions by engaging teachers in focused discussions about the goals of their instruction. There are also informal reports, however, from Great Britain that teachers are in near mutiny because of the inordinate time demands of the new assessment and curriculum system. In the U.S. we will have peculiar problems to resolve regarding the participation of individual teachers. While teachers evaluating their own students would make the most sense from an instructional point of view, and therefore would most justify use of the teachers' time, the distrust of teachers that motivates accountability demands in the first place would make it necessary to have teachers judging students other than their own.

Conclusion

There is a world of difference between the kinds of authentic, curriculum-driven examinations that the National Education Goals Panel

recommended be developed by the end of the decade and low-cost every-pupil tests that could possibly be developed in the short-term in response to the President's proposal for tests by 1994. Research evidence on the effects of traditional standardized tests when used as high-stakes accountability instruments is strikingly negative. It would not be far fetched to say that testing in the past decade has actually reduced the quality of instruction for many students rather than improving education. If tests are developed in advance of curriculum change, without teacher training, and imposed externally, with factory-like ideas of how to create scores, then it is likely that new tests will have many of the same pernicious effects as old tests.

References

Bauer, S., Mathison, S., Merriam, E., & Toms, K. (1990, April). *Controlling curricular change through state-mandated testing: Teachers' views and perceptions.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Borko, H., & Eisenhart, M. (1986). Students' conceptions of reading and their reading experiences in school. *The Elementary School Journal, 86,* 589-611.

Bracey, G.W. (1987). Measurement-driven instruction: Catchy phrase, dangerous practice. *Phi Delta Kappan, 68,* 683-686.

California State Department of Education. (1988). *Here they come ready or not: A report of the School Readiness Task Force.* Sacramento, CA: Author.

Cannell, J.J. (1987). *Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average.* Daniels, WV: Friends for Education.

Catterall, J. (1988). *Standards and school dropouts: A national study of the minimum competency test.* Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Cunningham, A.E. (1989). *Eeny, meeny, miny, moe: Testing policy and practice in early childhood.* Paper prepared for the National Commission on Testing and Public Policy, University of California, Berkeley.

Darling-Hammond, L. (1988). Accountability and teacher professionalism. *American Educator, 12,* 8-13, 38-43.

Darling-Hammond, L., & Wise, A.E. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85,* 315-336.

Debra P. v. Turlington, 644 F. 2d 397, 5th Cir. 1981: 564 F. Supp.177 (M.D. Fla. 1983).

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18,* 27-32.

Hatch, J.A., & Freeman, E.B. (1988). Who's pushing whom? Stress and kindergarten. *Phi Delta Kappan, 69,* 145-147.

Koretz, D.M., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991, April). *The effects of high-stakes testing on achievement: Preliminary findings about*

generalization across tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Linn, R.L., Graue, M.E., & Sanders, N.M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9,* 5-14.

McNeil, L.M. (1988). Contradictions of control, part 3: Contradictions of reform. *Phi Delta Kappan, 69,* 478-485.

A Nation at Risk. (1983). Washington, DC: The National Commission on Excellence in Education.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

National Education Goals Panel. (1991, March). *Measuring progress toward the national educational goals: Potential indicators and measurement strategies.* Washington, DC: Author.

Popham, W.J., Cruse, K.L., Rankin, S.C., Sandifer, P.D., & Williams, R.L. (1985). Measurement-driven instruction: It's on the road. *Phi Delta Kappan, 66,* 628-634.

Resnick, L.B., & Resnick, D.P. (in press). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford & M.C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction.* Boston: Kluwer Academic Publishers.

Rottenberg, C., & Smith, M.L. (1990, April). *Unintended effects of external testing in elementary schools.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Schoenfeld, A.H. (in press). On mathematics as sense-making: An informal attack on the unfortunate divorce of formal and informal mathematics. In D.N. Perkins, J. Segal, & J. Voss (Eds.), *Informal reasoning and education.* Hillsdale, NJ: Erlbaum.

Shepard, L.A. (1988, April). *Should instruction be measurement driven: A debate.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Shepard, L.A. (1989). A review of research on kindergarten retention. In L.A. Shepard & M.L. Smith (Eds.), *Flunking grades: Research & policies on retention.* London: Falmer Press.

Shepard, L.A. (1989). Why we need better assessments. *Educational Leadership, 46,* 4-9.

Shepard, L.A. (1991, April). *Effects of high-stakes testing on instruction.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Shepard, L.A., & Smith, M.L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. *The Elementary School Journal, 89*, 135-145.

Shepard, L.A., & Smith, M.L. (Eds.). (1989). *Flunking grades: Research and policies on retention.* London: Falmer Press.

Tyson-Bernstein, H. (1988). America's textbook fiasco. *American Educator, 12*, 20-27, 39.

U.S. Department of Education. (1991). *America 2000.* Washington, DC: Author.

Walberg, H.J. (1970). Professional role discontinuities in educational careers. *Review of Educational Research, 40*, 409-420.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.